Importing the required modules

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from seaborn.rcmod import palettes
import datetime as dt
import warnings
warnings.filterwarnings('ignore')
```

Importing the dataset

```
data=pd.read_csv(r'/content/hotel_booking.csv')
```

Data Exploration

```
data.head(2)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_dat
0	Resort Hotel	0	342	2015	July	
1	Resort Hotel	0	737	2015	July	

```
data.tail(2)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arriva
119388	City Hotel	0	109	2017	August	
119389	City Hotel	0	205	2017	August	

```
data.shape
```

```
(119390, 36)
```

```
data.columns
```

```
Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   hotel                           119390 non-null  object 
 1   is_canceled                     119390 non-null  int64  
 2   lead_time                       119390 non-null  int64  
 3   arrival_date_year               119390 non-null  int64  
 4   arrival_date_month              119390 non-null  object 
 5   arrival_date_week_number        119390 non-null  int64  
 6   arrival_date_day_of_month       119390 non-null  int64  
 7   stays_in_weekend_nights         119390 non-null  int64  
 8   stays_in_week_nights            119390 non-null  int64  
 9   adults                          119390 non-null  int64  
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64  
 12  meal                            119390 non-null  object 
 13  country                         118902 non-null  object 
 14  market_segment                  119390 non-null  object 
 15  distribution_channel            119390 non-null  object 
 16  is_repeated_guest               119390 non-null  int64  
 17  previous_cancellations          119390 non-null  int64  
 18  previous_bookings_not_canceled  119390 non-null  int64  
 19  reserved_room_type              119390 non-null  object 
 20  assigned_room_type              119390 non-null  object 
 21  booking_changes                 119390 non-null  int64  
 22  deposit_type                    119390 non-null  object 
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64  
 26  customer_type                   119390 non-null  object 
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64  
 29  total_of_special_requests       119390 non-null  int64  
 30  reservation_status              119390 non-null  object 
 31  reservation_status_date         119390 non-null  object 
 32  name                            119390 non-null  object 
 33  email                           119390 non-null  object 
 34  phone-number                    119390 non-null  object 
 35  credit_card                     119390 non-null  object 
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

```
data.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	a
count	119390.000000	119390.000000	119390.000000	119390.000000	
mean	0.370416	104.011416	2016.156554	27.165173	
std	0.482918	106.863097	0.707476	13.605138	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	160.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

```
data.duplicated().unique()
```

```
array([False])
```

```
data.isnull().sum()
```

```
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date_year               0
arrival_date_month              0
arrival_date_week_number        0
arrival_date_day_of_month       0
stays_in_weekend_nights         0
stays_in_week_nights            0
adults                          0
children                        4
babies                          0
meal                            0
```

```
country                           488
market_segment                      0
distribution_channel                0
is_repeated_guest                   0
previous_cancellations              0
previous_bookings_not_canceled      0
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
agent                           16340
company                        112593
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
reservation_status_date             0
name                                0
email                               0
phone-number                        0
credit_card                         0
dtype: int64
```

```
data.describe(include = 'object')
```

	hotel	arrival_date_month	meal	country	market_segment	distribution_
count	119390	119390	119390	118902	119390	
unique	2	12	5	177	8	
top	City	August	BB	PRT	Online TA	

Data Cleaning

Converting 'reservation_status_date' to date time format

```
data['reservation_status_date']=pd.to_datetime(data['reservation_status_date'])
```

```
data['reservation_status_date']
```

```
0        2015-07-01
1        2015-07-01
2        2015-07-02
3        2015-07-02
4        2015-07-03
            ...
119385   2017-09-06
119386   2017-09-07
119387   2017-09-07
119388   2017-09-07
119389   2017-09-07
Name: reservation_status_date, Length: 119390, dtype: datetime64[ns]
```

Removing the coloumns with most null values

```
data.drop(['company','agent'], axis=1, inplace= True)
```

Removing the null values

```
data.dropna(inplace= True)
```

```
data =data[data['adr']<50000]
```

```
data['month']=data['reservation_status_date'].dt.month
```

Data Analysis with Visualization

Checking the cleaned data

```
data.head(2)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_d
0	Resort Hotel	0	342	2015	July	
1	Resort Hotel	0	737	2015	July	

```
data.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	a
count	118898.000000	118898.000000	118898.000000	118898.000000	
mean	0.371352	104.311435	2016.157656	27.166555	
std	0.483168	106.903309	0.707459	13.589971	
min	0.000000	0.000000	2015.000000	1.000000	
25%	0.000000	18.000000	2016.000000	16.000000	
50%	0.000000	69.000000	2016.000000	28.000000	
75%	1.000000	161.000000	2017.000000	38.000000	
max	1.000000	737.000000	2017.000000	53.000000	

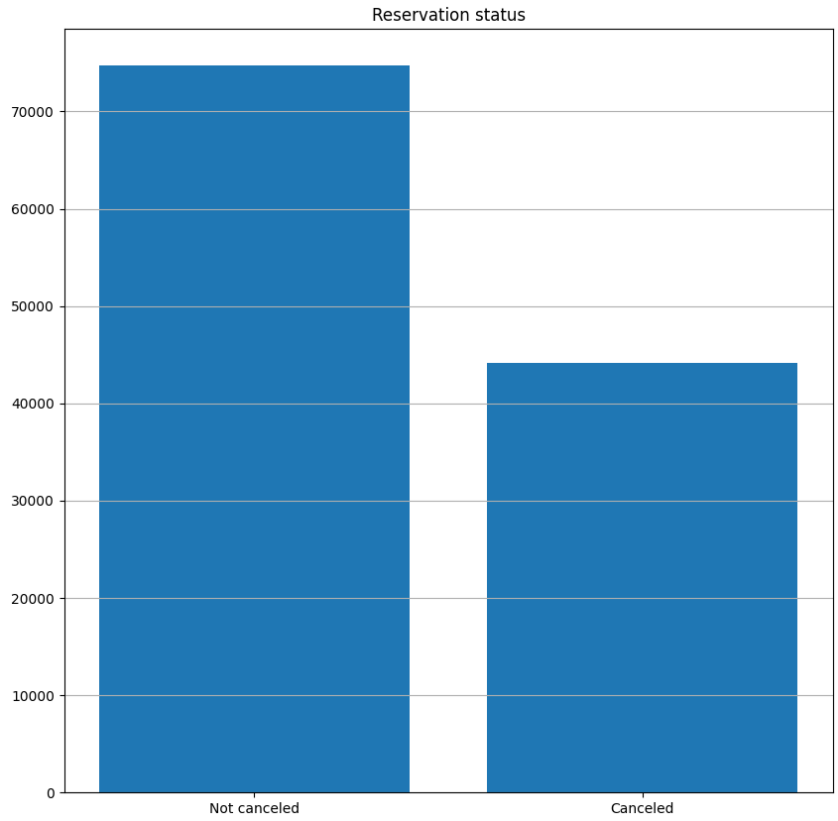
```
data.isnull().sum()
```

```
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date_year               0
arrival_date_month              0
arrival_date_week_number        0
arrival_date_day_of_month       0
stays_in_weekend_nights         0
stays_in_week_nights            0
adults                          0
children                        0
babies                          0
meal                            0
country                         0
market_segment                  0
distribution_channel            0
is_repeated_guest               0
previous_cancellations          0
previous_bookings_not_canceled  0
reserved_room_type              0
assigned_room_type              0
booking_changes                 0
deposit_type                    0
days_in_waiting_list            0
customer_type                   0
adr                             0
required_car_parking_spaces     0
total_of_special_requests       0
reservation_status              0
reservation_status_date         0
name                            0
email                           0
phone-number                    0
credit_card                     0
month                           0
dtype: int64
```

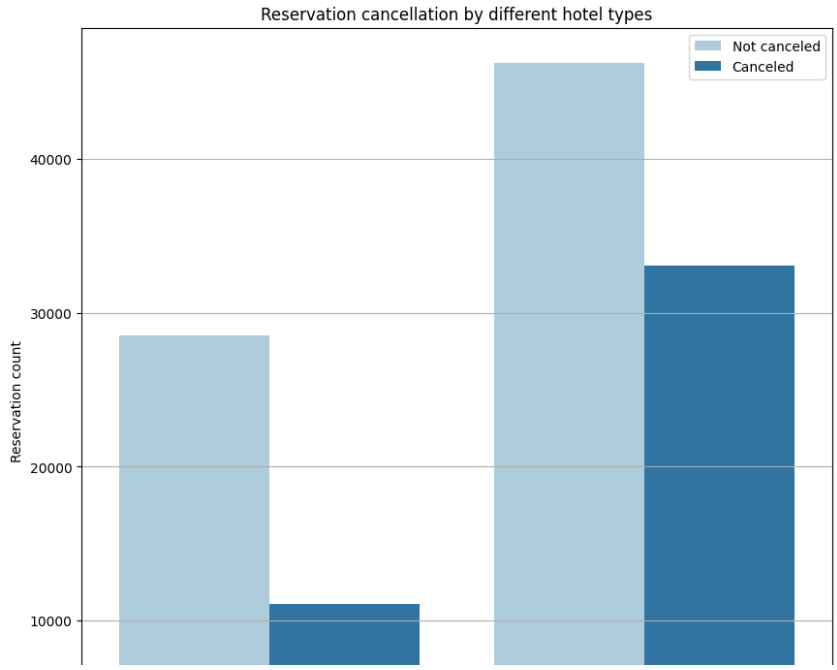
Calculating the Cancellation percentage and Visualizing

```
percentage_cancelled = data['is_canceled'].value_counts(normalize = True)
```

```
plt.figure(figsize = (10,10))
plt.title('Reservation status')
plt.bar(['Not canceled','Canceled'],data['is_canceled'].value_counts())
plt.grid(visible = bool, axis = 'y')
plt.show()
```




```
plt.figure(figsize=(10,10))
ax1=sns.countplot(x=data['hotel'],hue=data['is_canceled'],data=data,palette='Paired')
legend_labels=ax1.get_legend_handles_labels()
plt.title('Reservation cancellation by different hotel types')
plt.xlabel('Hotel type')
plt.ylabel('Reservation count')
plt.legend(['Not canceled','Canceled'])
plt.grid(visible = bool, axis = 'y')
plt.show()
```



Checking the cancellation percentage of Resort and City hotels

```
resort_hotel=data[data['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize= True)
```

```
0    0.72025
1    0.27975
Name: is_canceled, dtype: float64
```

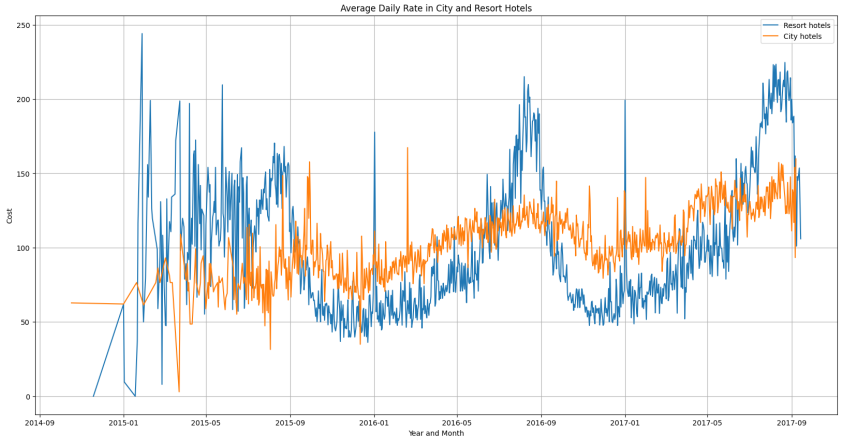
```
city_hotel=data[data['hotel']=='City Hotel']
city_hotel['is_canceled'].value_counts(normalize= True)
```

```
0    0.582911
1    0.417089
Name: is_canceled, dtype: float64
```

Creating a visualization for average daily rate

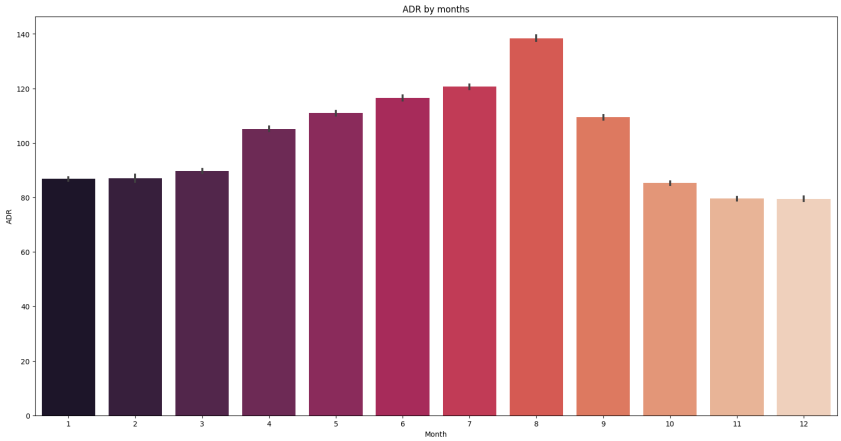
```
resort_hotel=resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel=city_hotel.groupby('reservation_status_date')[['adr']].mean()
```

```
plt.figure(figsize=(20,10))
plt.plot(resort_hotel.index,resort_hotel['adr'],label='Resort Hotels')
plt.plot(city_hotel.index,city_hotel['adr'],label='City Hotels')
legend_labels=ax1.get_legend_handles_labels()
plt.title('Average Daily Rate in City and Resort Hotels')
plt.xlabel('Year and Month')
plt.ylabel('Cost')
plt.legend(['Resort hotels','City hotels'])
plt.grid(visible = bool, axis = 'both')
plt.show()
```




```
plt.figure(figsize=(20,10))
ax1=sns.countplot(x=data['month'],hue=data['is_canceled'],data=data,palette='mako')
legend_labels=ax1.get_legend_handles_labels()
plt.title('Reservation cancellation by months')
plt.xlabel('Month')
plt.ylabel('Cancelation count')
plt.legend(['Not canceled','Canceled'])
plt.show()
```

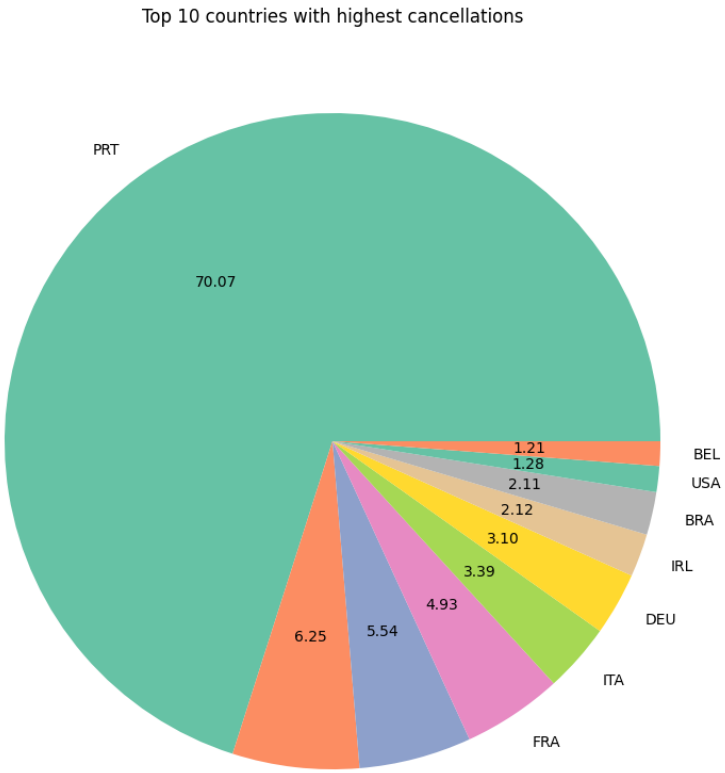

```
plt.figure(figsize=(20,10))
sns.barplot(x=data['month'],y=data['adr'], data = data[data['is_canceled'] == 1].groupby('month')[['adr']].sum().reset_i
plt.title('ADR by months')
plt.xlabel('Month')
plt.ylabel('ADR')
plt.show()
```



Analysis based on top 10 countries

```
canceled_country_data= data[data['is_canceled']==1]
top_10_countries = canceled_country_data['country'].value_counts()[:10]
```

```
plt.figure(figsize=(10,10))
colors = sns.color_palette('Set2')[0:10]
plt.pie(top_10_countries,autopct='%.2f',labels=top_10_countries.index, colors=colors)
plt.title('Top 10 countries with highest cancellations')
plt.show()
```




```
not_canceled_country_data= data[data['is_canceled']==0]
bottom_10_countries = not_canceled_country_data['country'].value_counts()[:10]


plt.figure(figsize=(10,10))
colors = sns.color_palette('husl')[0:10]
plt.pie(bottom_10_countries,autopct='%.2f',labels=top_10_countries.index, colors=colors)
plt.title('Bottom 10 countries with lowest cancellations')
plt.show()
```

Bottom 10 countries with lowest cancellations

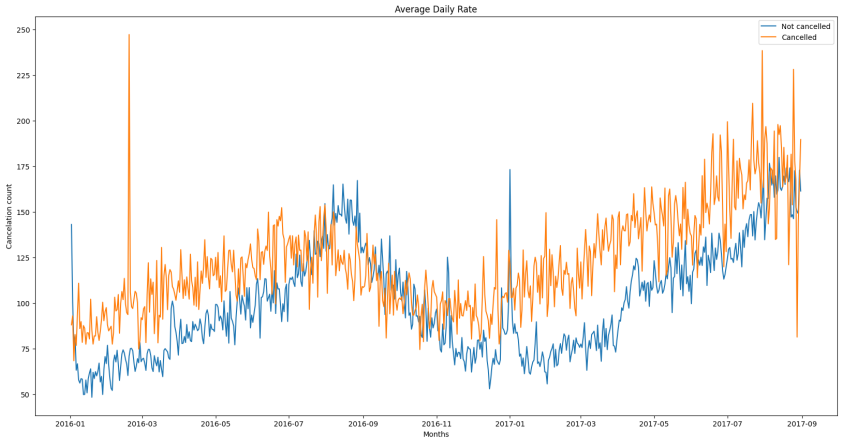
Analysing on the basis of price

```
canceled_data_adr = canceled_country_data.groupby('reservation_status_date')[['adr']].mean()
canceled_data_adr.reset_index(inplace=True)
canceled_data_adr.sort_values('reservation_status_date', inplace = True)
canceled_data_adr = canceled_data_adr [(canceled_data_adr['reservation_status_date']> '2016') & (canceled_data_adr['rese
```

```
not_canceled_data_adr = not_canceled_country_data.groupby('reservation_status_date')[['adr']].mean()
not_canceled_data_adr.reset_index(inplace=True)
not_canceled_data_adr.sort_values('reservation_status_date', inplace=True)
not_canceled_data_adr = not_canceled_data_adr[(not_canceled_data_adr['reservation_status_date']> '2016') & (not_canceled
```

2.58

```
plt.figure(figsize=(20,10))
plt.title('Average Daily Rate')
plt.plot(not_canceled_data_adr['reservation_status_date'], not_canceled_data_adr['adr'], label = 'Not cancelled')
plt.plot(canceled_data_adr['reservation_status_date'], canceled_data_adr['adr'], label = 'Cancelled')
plt.xlabel('Months')
plt.ylabel('Cancelation count')
plt.legend()
plt.show()
```



✓ 1s completed at 5:11 PM